# Definizione di una procedura di codifica delle domande aperte basata sui modelli delle indagini internazionali

**Giorgio Asquini**

*«Sapienza» Università di Roma, Dipartimento di Psicologia dei Processi di Sviluppo e Socializzazione*

doi: 10.7358/ecps-2014-010-asqu giorgio.asquini@uniroma1.it

## DEFINITION OF A CODING PROCEDURE OF OPEN-ENDED QUESTIONS BASED ON THE MODELS OF INTERNATIONAL STUDIES

### Abstract

*The paper presents an in-depth study within the «Problem-solving and geographical skills» project funded by Sapienza University of Rome in 2011. The main research instrument consists of open-ended items which was used as a basis for designing a coding procedure of student responses in order to maximise coding and coder control and reliability. Bearing in mind that when analyzing open answers, the margins of the assessor's subjective interpretation are certainly greater than with classic structured items, the adoption of strict encoding and coding control procedures allows us to steer this type of semi-structured questions toward objectivity, thereby significantly reducing errors of interpretation. The main reference is the procedure used in the OECD-PISA for open-ended questions, but also the assessment of written tests established in the IEA study on «Written composition». The aim is to improve the quality of the dataset of the study, with the least possible burden on resources. The results confirmed the effectiveness of the procedure in terms of reliability of coding, with the estimation of a low-level error. It was also possible to provide timely feedback to each coder, thereby enabling an improvement in coding ability.*

*Keywords:* Assessment, Coding, Open-ended items, Problem-solving, Reliability.

## 1. Introduzione: perché le domande aperte

Una delle novità più rilevanti dell'indagine OCSE-PISA, fin dal suo primo ciclo del 2000, è stato il massiccio uso di *item* a risposta aperta per la stima della *literacy* degli studenti quindicenni. La decisione di utilizzare questo tipo di strumento fu motivata con la necessità di indagare in modo più approfondito i livelli più elevati di abilità (OECD, 1999, p. 52), laddove i quesiti a risposta chiusa non riescono a fornire informazioni precise. Nello stesso quadro di riferimento del primo ciclo di PISA si preconizza che l'uso di domande aperte è destinato a crescere, una volta risolti alcuni problemi metodologici: «The extent to which this type of exercise will be used will depend on how robust the methodology proves to be in the field trial and how consistent a form of marking can be developed» (OECD, 1999, p. 14). Nonostante le cautele, la percentuale di domande aperte utilizzate risulta molto alta fin da PISA 2000: sono il 45% del totale degli *item* della *literacy* in Lettura e per la sottoscala più complessa (Riflessione e Valutazione) raggiungono addirittura il 65%: due domande su tre richiedono allo studente di scrivere per esteso la risposta (OECD, 1999, p. 37). La progressiva crescita pronosticata nel *framework* di PISA 2000 trova conferma nell'ultima rilevazione in Lettura di PISA 2012, in cui le domande aperte sono più della metà del totale [1].

L'impatto degli *item* a risposta aperta sui risultati degli studenti è quindi assai rilevante, e per il nostro paese costituisce uno de motivi principali della *performance* in Lettura, sempre inferiore alla media OCSE (Asquini & Corsini, 2010, p. 200).

La grande rilevanza data agli esiti di PISA comporta inevitabilmente una forte responsabilità per l'OCSE, i cui modelli di indagine diventano riferimenti per la valutazione in tutti i paesi membri, per quanto riguarda tutti gli aspetti, dalla definizione degli ambiti e delle scale, alle modalità di analisi dei dati, alla costruzione degli strumenti (McGaw, 2008, p. 229). Con un certo ritardo l'indicazione di puntare sulle domande aperte sta prendendo piede anche nel nostro paese, visto che solo dal 2010 appaiono anche nelle rilevazioni nazionali Invalsi (INVALSI, 2010, p. 20) e la tendenza è orientata a un progressivo incremento del loro utilizzo (INVALSI, 2014, pp. 17-18), con la conseguente definizione di una serie di procedure, dalla costruzione dei quesiti alla modalità di codifica, fino alla definizione dei punteggi, che si fondano soprattutto sull'esperienza di PISA. D'altra parte l'uso di questo tipo di strumento, almeno per rilevazioni su larga scala, ha faticato ad imporsi,

[1] Non è stato ancora pubblicato il *Technical report* di PISA 2012, ma le procedure di codifica svolte permettono di anticipare che le domande aperte sono state 25 su 45, contando due volte una domanda aperta che prevedeva due livelli di risposta.

ma ormai risulta attestato in diversi contesti nazionali (Anderson & Morgan, 2008; Morris, 2011).

Risulta quindi rilevante una riflessione sul tema delle domande aperte, cercando di approfondirne potenzialità e problemi, con lo scopo di fornire indicazioni operative per il loro utilizzo anche in contesti più ristretti rispetto a quelli delle rilevazioni nazionali e internazionali, in modo da arricchire l'armamentario strumentale di chi opera direttamente nelle scuole.

## 2. I problemi delle domande aperte e le risposte di PISA

La scelta di puntare sulle domande aperte in PISA naturalmente si è fondata sulle principali evidenze teoriche circa questo tipo di strumento (Ward, Dupree, & Carlson, 1987; Ackerman & Smith, 1988; Bennett & Ward, 1993). Il principale problema dell'utilizzo delle domande aperte riguarda la trasformazione della risposta dello studente in un codice che ne attesti la correttezza. Anche se si stanno sperimentando modalità di analisi automatica dei testi liberi prodotti dagli studenti (Bolasco, 2010), è evidente che l'intervento umano di decodifica della risposta dello studente risulta fondamentale. È proprio questa interpretazione che risulta essere la guida essenziale per gli altri due problemi delle domande aperte: la costruzione dell'*item* e il trattamento dei punteggi derivati dalle codifiche.

Nella definizione dell'*item* la difficoltà principale è costituita dalla formulazione della domanda, dalla cui lettura deve risultare chiaro e non interpretabile il compito cognitivo richiesto per rispondere. Strettamente collegata è l'indicazione specifica sulle modalità di risposta, con una consegna precisa su come rispondere, compreso il formato e lo spazio di risposta. Lo studente deve quindi aver chiaro cosa deve fare e come deve rispondere. In questa fase di elaborazione dell'*item* si procede anche alla stesura di una guida alla codifica delle risposte aperte (*Coding guides*, OECD, 2012, p. 39), in cui si definiscono le modalità di risposta corretta e non corretta (quindi anche i possibili errori in cui può incorrere lo studente), complete di esempi di risposta.

Nell'esperienza di PISA i quesiti e le guide sono messe alla prova nel *field trial* che precede ogni studio principale, verificando se le risposte degli studenti soddisfano le intenzioni di verifica della *literacy* ipotizzate dagli estensori della prova: tanto più le risposte si allontanano dai processi cognitivi richiesti, tanto meno la domanda aperta risulta capace di stimare. La verifica è quindi affidata in primo luogo ai codificatori esperti e solo successivamente a procedure statistiche di *item* analisi. Le domande che riescono a

superare questa fase di verifica esperta possono essere affinate (in particolare per quanto riguarda gli aspetti grafici e le modalità di risposta), ma soprattutto vengono integrate le istruzioni di codifica, puntualizzando meglio le risposte accettabili e non accettabili e ampliando il quadro degli esempi di risposta possibili. In questo modo i codificatori esperti che dovranno operare nello studio principale potranno disporre di uno strumento di supporto alla codifica molto più efficiente e orientato all'oggettività, che resta sempre il problema principale delle domande aperte.

Nella fase di verifica viene testata anche la possibilità che una domanda riesca a distinguere diversi livelli di accettabilità della risposta, con un punteggio pieno e uno (o più di uno) parziale. Fin dalla stesura del quesito infatti può risultare evidente che siano possibili risposte non complete, ma che vanno nella direzione del compito cognitivo richiesto per rispondere, soprattutto per le domande di maggiore complessità. Il *field trial* permette di confermare o modificare questa distinzione, e in questo caso l'incrocio delle opinioni dei codificatori con la verifica statistica risulta decisiva, perché può succedere che l'ipotesi di punteggio parziale formulata dagli estensori della domanda, risulti teorica, in quanto non attestata dalle risposte effettive degli studenti. In PISA le domande che prevedono diversi livelli di punteggio sono in progressiva diminuzione [2].

Nonostante la rigorosità della procedura le domande aperte restano pur sempre uno strumento semistrutturato (Domenici, 1996), per cui alcune risposte degli studenti possono sfuggire alle buone intenzioni di codifica (cioè non trovare riscontro nella guida) o anche essere equivocate dal codificatore. L'equivoco è spesso fondato sugli aspetti linguistici della risposta (grammaticali, ma anche grafici), che possono interferire nei processi di decodifica di quanto lo studente ha scritto. Ma l'equivoco può anche derivare da interferenze di lettura (stanchezza, distrazione), possibile anche per un codificatore esperto. Bisogna quindi mettere in conto un possibile livello di errore, sicuramente maggiore rispetto a quello dei quesiti strutturati. Per ridurre al minimo i possibili errori PISA ha predisposto diverse procedure di assistenza alla codifica e di controllo dei codificatori. Prima di illustrarle nel dettaglio diciamo subito che si tratta di procedure impegnative dal punto di vista economico, che pongono seriamente il tema del rapporto costi/benefici circa l'uso dello strumento (Toch, 2006, p. 14), aggiungendo nei costi anche il maggior tempo necessario al trattamento delle risposte e il conseguente allungamento dei tempi di analisi e restituzione dei risultati. D'altra parte chi le ha utilizzate conferma che il quadro informativo sulle capacità degli

[2] In PISA 2012 una sola domanda aperta su 24 ha previsto anche un livello di punteggio parziale.

studenti risulta molto più ampio, articolato, motivato e utile per la didattica, per cui non si tratta di costi inutili, semmai il problema è relativo alla sostenibilità dei costi:

> «Constructed-response questions give you more measurement depth», says John Olson, director of psychometric and research services at Harcourt Assessment and director of assessment at the Council of Chief State School Officers from 1998 to 2003. «They give you a better sense of what students can do. And as a result, teachers get more out of them». (Toch, 2006, p. 16)

Ma vediamo più nel dettaglio le procedure che PISA utilizza per garantire la qualità delle codifiche delle domande aperte[3]. Tutte le procedure sono realizzate in collaborazione fra l'ambito nazionale e l'ente operativo che coordina la ricerca a livello internazionale[4], poiché uno dei problemi principali di una ricerca internazionale riguarda l'uniformità delle procedure in tutti i paesi partecipanti.

In primo luogo si svolge la selezione dei codificatori, realizzata a livello nazionale, ma basata su un materiale di formazione internazionale (*Coder recruitment kit*, OECD, 2012, p. 39) per verificare la capacità dell'aspirante codificatore di adeguarsi alle modalità di lavoro e alle indicazioni della guida alla codifica.

Viene stabilita una gerarchia nel gruppo di lavoro, con un *supervisor* e alcuni *table leaders* (entrambe questo figure hanno maturato una lunga esperienza di codifica e partecipato agli incontri internazionali di formazione dei codficatori), che devono fornire un costante supporto ai codificatori, cercando di risolvere i casi dubbi nonché vigilare sulle condizioni di lavoro.

L'OCSE fornisce indicazioni ai responsabili della codifica in ordine agli spazi e ai tempi di lavoro, in particolare scoraggiando sessioni troppo lunghe e prevedendo opportune pause. Inoltre la scansione delle operazioni è definita nei documenti internazionali e deve essere garantita dai gruppi di lavoro nazionali. Si procede secondo un disegno progressivo di distribuzione delle domande ai codificatori (OECD, 2012, p. 107) che considera i gruppi

[3] Tutte le procedure sono illustrate nel *Technical report* che accompagna ogni ciclo di PISA. L'ultimo pubblicato è quello relativo al ciclo 2009 (OECD, 2012), ma le procedure risultano sostanzialmente le stesse in tutti i cicli. In particolare si fa riferimento ai capp. 2 («Test design e test development»), 6 («Field operations») e 13 («Coding reliability study»). Il documento specifico sulle codifiche utilizzato dai National Project Manager (*Procedures for coding constructed-response items MS09*, citato in OECD, 2012, p. 102) contiene riferimenti ai materiali di ricerca, per cui è ovviamente riservato per garantire la ripetibilità degli *item* nei cicli successivi.

[4] Fino al ciclo 2012 l'indagine PISA è stata coordinata, per conto dell'OCSE, dall'Australian Council for Educational Research (ACER).

(*clusters*) di domande e i relativi fascicoli in cui sono contenuti, completando tutte le risposte relative a una domanda prima di passare alla domanda successiva (i codificatori quindi devono procedere sincronicamente, senza salti in avanti o ritardi). Questo permette di non sovrapporre problemi relativi a domande diverse, ma soprattutto permette ai coordinatori di seguire con maggior precisione i lavori, poiché al termine di ogni giornata di lavoro i *table leaders* effettuano un controllo a campione sulle codifiche assegnate, correggendo eventuali errori e richiamando i codificatori che abbiano mostrato maggiori oscillazioni (nella pratica ciò accade se si rivela un tasso di errore superiore al 10%). Il raccordo puntuale fra lavoro svolto e verifica effettuata risulta essere molto motivante per i codificatori.

Nonostante l'accuratezza delle guide di codifica è possibile che un codificatore abbia dei dubbi di fronte ad una risposta particolarmente originale, e se il dubbio non viene risolto dal confronto con un *table leader* o con il *supervisor*, è possibile rivolgersi a un *Coder Query Service* internazionale (OECD, 2012, p. 45) che raccoglie tutti i dubbi e cerca di fornire indicazioni dirimenti praticamente in tempo reale. Inoltre le segnalazioni provenienti dai diversi paesi vengono aggiunte in appositi elenchi di esempi commentati che integrano ulteriormente le guide di codifica. L'esperienza maturata nei gruppi di codifica da PISA 2000 a oggi permette di dire che anche alla terza-quarta riproposizione di un *item* possono emergere esempi di risposta originali, su domande in cui sono state già fornite decine di migliaia di risposte.

Un ulteriore passaggio nelle procedure di codifica è finalizzato alla verifica dell'affidabilità dei singoli codificatori e di conseguenza all'oggettività delle codifiche a livello nazionale. Si tratta della fase finale delle codifiche multiple (OECD, 2012, p. 109), in cui i codificatori vengono divisi in gruppi di 4, a ogni gruppo vengono assegnati 100 fascicoli e ogni codificatore codifica singolarmente, senza contatti con i colleghi, tutte le domande aperte contenute nei 100 fascicoli. In questo modo è possibile confrontare le 4 codifiche sullo stesso *item* e verificare il grado di accordo dei codificatori su un *set* di domande comuni. Considerando questo *set* di codifiche multiple è possibile calcolare un *Reliability index* (OECD, 2012, p. 235) per ogni codificatore e per ogni domanda, e raggruppando i dati verificare il grado di accordo delle codifiche per tutto il gruppo dei codificatori e per tutte le domande aperte. Da notare che non vengono effettuati riscontri sull'eventuale errore, cioè si stima semplicemente che uno o più codificatori abbiano sbagliato. Il livello di accordo è quindi lo stimatore diretto dell'affidabilità, intesa come contenimento dell'errore. Per il definitivo passaggio alla verifica dell'oggettività è però necessario un ulteriore fase, in quanto 4 codificatori potrebbero anche andare perfettamente d'accordo, ma nell'errore. Quindi un campione di fascicoli già codificati viene inviato all'ACER, ricodificato da

esperti internazionali e queste nuove codifiche vengono confrontate con le codifiche fornite dai codificatori nazionali, arrivando a una percentuale definitiva di accordo che conferma, o meno, la presenza del dato nelle analisi dei risultati. Di passaggio notiamo che i codificatori italiani in PISA 2009 hanno avuto complessivamente ottime valutazioni, non risultando ne troppo *harsh* ne troppo *lenient* (OECD, 2012, p. 244).

Le procedure illustrate presentano un alto grado di complessità, dipendente in larga misura dalle dimensioni dello studio, basato su uno strumento complesso (molte domande) e su un enorme campione internazionale. È possibile costruire una serie di procedure semplificate e adatte a uno studio di dimensioni più limitate mantenendo gli stessi standard di qualità?

## 3. L'indagine sul «problem solving» geografico

Il progetto *Problem solving e abilità geografiche* è stato realizzato tra il 2012 e il 2014, grazie al Finanziamento di Ateneo 2011 dell'Università «Sapienza» di Roma. Si tratta di uno studio correlazionale indirizzato a identificare i fattori, scolastici e personali, che possono influenzare le capacità di risoluzione di situazioni problematiche in ambito geografico, rilevate attraverso una prova specifica. Il campione di indagine è costituito da 1.235 studenti di 59 classi terze di scuola secondaria di I grado della provincia di Roma. Per la definizione del campione è stato sfruttato un precedente lavoro di ricerca che aveva indagato il fenomeno del *Summer Learning Loss* (Sabella, 2014) nel passaggio fra la prima e la seconda classe, per cui di circa 700 studenti sono disponibili anche i dati storici di profitto in prove di competenza della lettura. La decisione di occuparsi del *problem solving* è ancora una volta legata a PISA. Dopo il 2003, anno in cui il *problem solving* è stato rilevato come ambito specifico, gli stimoli venuti dall'indagine internazionale hanno trovato scarsa eco in ambito nazionale, limitandosi all'originale collegamento con l'ambito matematico. La riflessione scientifica avviata già dalla metà del secolo scorso (Polya, 1945) e la riflessione politico-scolastico più recente (ricordiamo il progetto *De.Se.Co.*, Ryjchen & Salganik, 2000) avevano però già identificato nella capacità di risolvere qualsiasi tipo di problema una delle più importanti abilità trasversali che l'istruzione dovrebbe promuovere. L'applicazione di questa abilità a un contesto geografico è risultata particolarmente stimolante, vista anche la mancanza di studi specifici, escluse le prove orientate al movimento negli spazi presenti in PISA 2003 (OECD, 2004; Asquini, 2006). La costruzione della prova si è fondata sul modello di PISA per il *problem solving*, in particolare il *framework* del ciclo 2003 (OECD, 2003), poiché quello aggiornato del 2012

non era stato ancora pubblicato (OECD, 2013). I quadri di riferimento risultano comunque in stretta continuità fra di loro, con l'integrazione della componente di risoluzione computerizzata, poiché le prove di *problem solving* di PISA 2012 sono state somministrate esclusivamente al *computer*. Naturalmente sono state considerate anche indicazioni più aggiornate (Jonassen, 2011) e specificamente indirizzate alla popolazione di riferimento (Weber *et al.*, 2010), considerata interessante proprio perché negli anni della scuola secondaria di primo grado si verifica il passaggio dal pensiero concreto a quello formale, con la conseguente capacità di riconoscere e costruire modelli.

La prova è composta da 7 testi stimolo, per un totale di 23 *item*. Due testi (3 *item*, tutti a risposta aperta) sono ripresi dal pacchetto *Problem Solving* rilasciato di PISA 2003 [5], mentre 5 testi e 20 *item* sono originali. Di questi ultimi ben 15 sono aperti, e 8 di questi sono a risposta articolata (*Open-constructed response items*, OECD, 2004, p. 140), cioè richiedono alle studente di scrivere per esteso la risposta (o tracciare su una cartina/mappa percorsi complessi). Le altre 7 domande (10, se consideriamo le tre mutuate da PISA) richiedono invece o una risposta breve di tipo univoco (*Closed-constructed response items*, nel caso specifico un numero, che viene inserito direttamente senza codifica), o una risposta aperta breve (*Short response items*), che richiede comunque una codifica prima dell'inserimento dei dati. In tutto quindi gli *item* originali sottoposti a codifica sono 12, 8 a risposta articolata e 4 a risposta breve. Per questo, parallelamente alla costruzione dei quesiti, si è proceduto alla stesura della *Guida di codifica* [6], e dopo il *field trial* della prova la guida è stata modificata per renderla più aderente alle risposte reali degli studenti e integrata con un ampio numero di esempi di risposte, secondo il modello PISA visto sopra. I principali interventi sulla prova in seguito al *field trial* hanno riguardato soprattutto gli elementi grafici (disegni, cartine, mappe) a corredo delle domande, che sono stati ritoccati e migliorati anche in seguito a suggerimenti specifici tratti dagli studenti impegnati nel collaudo, che sono stati interpellati sulla chiarezza dei materiali proposti. Inoltre, considerata l'originalità della prova, l'esperienza del *field trial* è stata utilissima per la definizione dei tempi di somministrazione.

Il controllo delle risposte del *field trial* ha permesso anche di sciogliere i dubbi circa la necessità di distinguere diversi livelli di adeguatezza della

[5] Si tratta dei testi *Vacanze* e *Rete di trasporto*, che richiedono allo studente di risolvere situazione problematiche riguardanti la lettura di cartine e mappe (OECD, 2004, pp. 17 e 73). Questo permetterà, in sede di analisi dei dati, di avere un collegamento con i risultati della rilevazione 2003.

[6] Per le tre domande aperte di PISA è stata naturalmente utilizzata la guida di codifica originale.

risposta. Solo per tre domande originali, e due provenienti da PISA 2003, sono previsti il punteggio pieno e il punteggio parziale, quindi le difficoltà di codifica per queste 5 domande possono risultare maggiori.

Una volta approntata la prova definitiva è stata realizzata la somministrazione principale dei materiali [7] e si è posto il problema della codifica di circa 19.000 domande aperte (di cui due terzi a risposta articolata), con la necessità di costruire una procedura di codifica ispirata a PISA ma più gestibile per una ricerca locale e dagli obiettivi nettamente più limitati, privilegiando l'obiettivo di raggiungere un livello accettabile di oggettività delle codifiche e una incidenza marginale degli errori di codifica.

## 4. La nuova procedura di codifica delle domande aperte

La procedura cerca di coniugare qualità ed efficienza, considerando che si basa largamente su materiali originali e che non può contare su codificatori con esperienza in codifiche nazionali o internazionali. Inoltre le condizioni di lavoro non possono essere quelle di PISA, che prevedono sessioni in presenza a tempo pieno per l'intera durata delle codifiche. Un'attenzione specifica è quindi posta al controllo dei codificatori, oltre naturalmente alla qualità delle codifiche.

Sono stati selezionati 9 codificatori, per un carico ipotetico di circa 2.000 codifiche a testa, prevedendo un periodo lavorativo concentrato in 4-5 settimane. Pur non essendo specificamente esperti, tutti i codificatori (laureati o laureandi triennali e magistrali dei corsi pedagogici della «Sapienza») avevano maturato esperienze di somministrazione di prove, lettura di testi prodotti dagli studenti (prove scritte e questionari), inserimento dei dati, codifiche dei codici di professione (codifiche ISCO 2008). L'aspetto motivazionale è stato rinforzato con la possibilità di ottenere crediti formativi utili per il percorso di studi e prevedendo un piccolo compenso basato sui fondi di ricerca. Tuttavia è stata considerata rischiosa la sola verifica finale prevista dal modello PISA tramite le codifiche multiple; il rischio era quello di accorgersi *a posteriori* che uno o più codificatori deviassero dalle indicazioni per la codifica. Inoltre il lavoro riguardava prove originali, e la stessa *Guida per la codifica* poteva risultare non del tutto adeguata, vista la mancanza di riscontri

[7] Il lavoro sul campo è stato svolto nell'ambito dell'Esercitazione di ricerca *Costruzione di prove per la scuola dell'obbligo*, del Corso di Laurea triennale di Scienze dell'Educazione e della Formazione della «Sapienza», Università di Roma. La somministrazione principale si è svolta nei mesi di dicembre 2013 e gennaio 2014.

su grandi campioni (il punto di partenza iniziale del *field trial* era comunque limitato a poco più di 100 studenti).

Ricordando il modello di correzione adottato dall'indagine IEA *Written composition* [8], che prevedeva per ogni elaborato scritto una doppia correzione cieca da parte di due correttori distinti, si è deciso di dividere i 9 codificatori in due gruppi, A (5 membri) e B (4 membri), con l'obiettivo di far codificare ogni domanda a due codificatori diversi, in una sorta di *peer review* applicata alle risposte degli studenti. I due gruppi hanno operato separati e senza contatti fra di loro e ad entrambi è stato richiesto di codificare l'intero *set* di domande, con due sole differenze: il gruppo A ha codificato tutte le 18 domande aperte e inserito in un apposito *database* le risposte dei 5 *item* chiusi, il gruppo B non ha codificato le 3 domande aperte a risposta assolutamente univoca e non si è occupato dell'inserimento delle domande chiuse. Ogni componente del gruppo A ha quindi codificato circa 4.500 risposte e inserito 1.200 risposte chiuse, mente ogni componente del gruppo B ha codificato circa 5.500 risposte. La doppia codifica comporta quindi un aumento del carico di lavoro complessivo, ma permette di avere per ogni risposta un riscontro da analizzare. Inoltre non si rende più necessaria la fase di verifica finale della qualità dei codificatori, poiché il monitoraggio si svolge sull'intero *set* di prove. Nei fatti i tempi di lavoro previsti sono stati rispettati.

Per entrambi i gruppi è stata svolta una formazione specifica basata sulla guida di codifica e sugli esempi tratti dal *field trial*, con esercizi individuali di codifica dei fascicoli dello stesso *field trial*: ognuno ha dovuto codificare (su una scheda) 5 fascicoli e poi passarli al collega vicino, secondo uno schema di rotazione che ha permesso di confrontare e discutere circa 300 codifiche date da tutti sulle stesse domande. In questo modo è stato anche possibile cominciare a capire quali domande creassero maggiori problemi di interpretazione e integrare ulteriormente gli esempi di risposte nella *Guida per la codifica*.

Per ovvie esigenze organizzative il lavoro dei componenti del gruppo non poteva essere svolto in presenza, per cui ognuno ha codificato a casa, ma sono state fornite indicazioni specifiche sulle modalità di lavoro: svolgere il lavoro in un tempo il più possibile ristretto (le scadenze di consegna erano comunque molto strette), in modo da mantenere la concentrazione sul lavoro da svolgere; procedere domanda per domanda, resistendo alla tentazione di codificare insieme due o più domande presenti nella stessa pagina

[8] La procedura di doppia correzione degli elaborati scritti utilizzata nell'indagine IEA era finalizzata soprattutto ad attenuare gli effetti di valutazioni divergenti da parte di uno o più valutatori per gli elaborati collocabili sulle diverse soglie della scala a 5 punti utilizzata nell'indagine, ma di fatto è stata utilizzata anche per verificare il livello di accordo, consenso e scarto del gruppo di valutatori (Albertoni, 1988; Benvenuto, 1993).

del fascicolo; inserire le codifiche in un *database* appositamente predisposto e non tracciare segni di alcun tipo sul fascicolo, per permettere il lavoro di verifica da parte di un altro codificatore (o dei supervisori). Considerando l'estrema utilità della fase di confronto svolta durante la formazione è stata inoltre messa a disposizione di ogni gruppo una pagina della piattaforma e-learning *Moodle* della «Sapienza», per poter disporre dei materiali necessari (prova, guida, esempi, *database* per l'inserimento), per poter consegnare le codifiche completate e soprattutto per segnalare in un apposito *forum* ulteriori risposte critiche e possibili esempi da integrare nella guida. Di fatto questo *forum* richiama il *Coder Query Service* di PISA, il cui scopo principale è di migliorare la qualità delle codifiche, ma la possibilità di interagire con gli altri componenti del gruppo e con i supervisori si è rivelata molto utile anche per motivare ulteriormente i codificatori: il gruppo A ha segnalato complessivamente 121 esempi di risposte critiche, il gruppo B ne ha segnalati 112, con un buon livello di partecipazione di ogni gruppo ai dubbi dei singoli e una larga sovrapposizione fra i due gruppi circa gli esempi segnalati, segno che una parte consistente dei problemi è stata condivisa anche senza contatti diretti fra i due gruppi. Queste segnalazioni hanno permesso di ipotizzare una categoria di verifica intermedia tra l'accordo pieno (due codificatori danno la stessa codifica a una risposta) e lo scarto (le due codifiche sono diverse), la categoria del consenso, definizione ripresa dalla procedura IEA *Written composition*. In quell'indagine il consenso si aveva sulle prove al confine fra due valutazioni vicine (Benvenuto, 1993, p. 27), basandosi sul principio che il giudizio in alcuni casi possa oscillare per quelle risposte che si trovano al confine fra due livelli di codifica (0-1 o 1-2). Nel nostro caso l'oscillazione doveva essere segnalata, quindi per le risposte dubbie segnalate nel *forum* e per quelle simili, l'eventuale differenza di codifica viene categorizzata a parte, poiché deriva anche dai limiti della *Guida per la codifica* di prevedere tutti i casi di risposta possibili. Naturalmente l'ampiezza di questa categoria deve essere il più possibile limitata, e vedremo nei risultati differenze interessanti fra i singoli *item*.

Una serie di accortezze è stata utilizzata per la distribuzione dei fascicoli fra i codificatori. Per il gruppo A la distribuzione è stata fatta classe per classe, in modo che tutti i codificatori ricevessero fascicoli di tutte le classi in misura simile, questo per due motivi: evitare che, nonostante i controlli, eventuali deviazioni individuali incidessero su una o poche classi; evitare il rischio che un codificatore si ritrovi a leggere solo fascicoli di studenti dello stesso livello di abilità, possibile nel caso in cui il suo pacco di fascicoli sia composto da poche classi intere, magari della stessa scuola. Quando un codificatore del gruppo A restituiva il suo pacco, i fascicoli che lo componevano venivano distribuiti in 4 pacchi diversi, in modo che tutte le combinazioni possibili

fra codificatori A e B fossero equivalenti dal punto di vista numerico, cioè ognuno si combinasse nella stessa misura con tutti i colleghi dell'altro gruppo. Come vedremo più avanti questo ha permesso una maggiore precisione nell'individuazione degli scarti individuali rispetto alla guida di codifica.

Il lavoro si è svolto sui materiali originali, per cui i due gruppi hanno dovuto lavorare in successione e complessivamente tutta la procedura ha richiesto poco meno di tre mesi, al termine dei quali è stato possibile avviare la fase di verifica incrociata delle codifiche date distintamente dai due gruppi su tutto il *set* di domande.

## 5. Risultati

Per procedere all'analisi dei risultati riguardanti l'affidabilità delle codifiche è stato necessario un lavoro di ricomposizione di tutti i file consegnati, che si è rivelato agevole poiché tutti i codificatori hanno rispettato le consegne tecniche sull'inserimento delle codifiche. Attraverso alcune procedure di *query* con *Microsoft Excel* è stato possibile ricavare una mappa dettagliata alla singola risposta di tutti i confronti, ed è stato quindi possibile riaggregare i dati per *item*, per gruppi e per singoli codificatori. Nella Tabella 1 sono riepilogate le percentuali di Accordo, Consenso e Scarto rilevate sull'intero *set* di 12 domande aperte originali (non sono considerate quelle riprese da PISA). Il totale è scomposto anche per i 4 *item* a risposta breve (2.2, 2.3, 2.4 e 5.5) e gli 8 a risposta articolata.

Complessivamente per oltre il 91% degli *item* c'è stato un pieno accordo fra i due codificatori. La percentuale sale a quasi il 99% per le risposte brevi, sicuramente più facili da interpretare, e il relativo 1% di scarto costituisce l'errore occasionale gratuito dovuto anche a fattori materiali (come un inserimento sbagliato). Per le risposte articolate i dati sono peggiori, ma la percentuale di scarto resta comunque nel complesso sotto il 10%, che è la soglia accettabile prevista da PISA. Se però si va nello specifico dei singoli *item*, vediamo che per due di loro (entrambi a risposta articolata, 1.3 e 4.1) la percentuale di scarto supera il 10%, segno che queste domande pongono problemi specifici di codifica per un numero di casi che comincia ad essere rilevante, e di conseguenza i codificatori peggiorano il loro livello di accordo. Da segnalare anche l'anomalia di un *item* (a risposta articolata, 5.4) che ha creato molti problemi di codifica segnalati nei *forum* (confluiti quindi nella categoria del Consenso), anche se alla fine lo scarto effettivo per questo *item* è risultato inferiore a quello delle domande articolate, segno che le criticità della *Guida* sono state rilevate nel corso dei lavori, e non alla fine come per

i due *item* precedentemente segnalati. Per quanto riguarda gli *item* 4.1 e 5.4 bisogna ricordare che per le risposte era previsto anche il punteggio parziale, quindi i margini di dubbio e di errore sono oggettivamente maggiori (di fatto ci sono due soglie critiche, fra 0 e 1 e fra 1 e 2).

Nella Tabella 2 sono riaggregati i dati relativi all'Accordo distinti fra i due gruppi e per ogni codificatore. Come si vede è stato possibile ricostruire tutti gli accordi di coppia, per cui ogni codificatore ha un dato complessivo e uno specifico per ogni collega dell'altro gruppo. Per agevolare la lettura i due dati di Non accordo (Consenso e Scarto) sono stati uniti.

*Tabella 1. – Riepilogo dell'accordo fra i codificatori (%).*

| Item | Accordo | Consenso | Scarto | Totale |
|---|---|---|---|---|
| 1.1 | 87,04 | 5,99 | 6,96 | 100,00 |
| 1.3 | 82,91 | 5,02 | 12,06 | 100,00 |
| 1.4 | 90,85 | 2,67 | 6,48 | 100,00 |
| 2.2 | 99,51 | 0,00 | 0,49 | 100,00 |
| 2.3 | 98,95 | 0,00 | 1,05 | 100,00 |
| 2.4 | 98,87 | 0,08 | 1,05 | 100,00 |
| 4.1 | 83,56 | 6,23 | 10,20 | 100,00 |
| 4.3 | 93,12 | 2,91 | 3,97 | 100,00 |
| 5.3 | 94,82 | 2,35 | 2,83 | 100,00 |
| 5.4 | 78,38 | 15,79 | 5,83 | 100,00 |
| 5.5 | 98,06 | 0,32 | 1,62 | 100,00 |
| 7.1 | 88,26 | 6,88 | 4,86 | 100,00 |
| Totale | 91,19 | 4,02 | 4,78 | 100,00 |
| Breve | 98,85 | 0,10 | 1,05 | 100,00 |
| Articolata | 87,37 | 5,98 | 6,65 | 100,00 |

Nota: i due gruppi erano composti da 5 (Gruppo A) e 4 (Gruppo B) codificatori, per un totale di 20 coppie.

*Tabella 2. – Controllo dell'accordo fra i gruppi e fra i codificatori (%).*

| Gruppo A (5 codificatori) | Accordo | Non Accordo | Gruppo B (4 codificatori) | Accordo | Non Accordo |
|---|---|---|---|---|---|
| Fra Gruppi | 91,19 | 8,81 | Fra Gruppi | 91,19 | 8,81 |
| A1-Gruppo B | 92,80 | 7,20 | B1-Gruppo A | 90,78 | 9,22 |
| A1-B1 | 93,71 | 6,29 | B1-A1 | 93,71 | 6,29 |
| A1-B2 | 90,78 | 9,22 | B1-A2 | 92,75 | 7,25 |
| A1-B3 | 93,73 | 6,27 | B1-A3 | 90,51 | 9,49 |
| A1-B4 | 92,95 | 7,05 | B1-A4 | 88,48 | 11,52 |
| A2-Gruppo B | 92,69 | 7,31 | B1-A5 | 88,38 | 11,62 |
| A2-B1 | 92,75 | 7,25 | B2-Gruppo A | 90,85 | 9,15 |
| A2-B2 | 92,16 | 7,84 | B2-A1 | 90,78 | 9,22 |
| A2-B3 | 92,35 | 7,65 | B2-A2 | 92,16 | 7,84 |
| A2-B4 | 93,54 | 6,46 | B2-A3 | 91,62 | 8,38 |
| A3-Gruppo B | 90,90 | 9,10 | B2-A4 | 90,48 | 9,52 |
| A3-B1 | 90,51 | 9,49 | B2-A5 | 89,22 | 10,78 |
| A3-B2 | 91,62 | 8,38 | B3-Gruppo A | 91,23 | 8,77 |
| A3-B3 | 89,90 | 10,10 | B3-A1 | 93,73 | 6,27 |
| A3-B4 | 91,57 | 8,43 | B3-A2 | 92,35 | 7,65 |
| A4-Gruppo B | 89,95 | 10,05 | B3-A3 | 89,90 | 10,10 |
| A4-B1 | 88,48 | 11,52 | B3-A4 | 90,48 | 9,52 |
| A4-B2 | 90,48 | 9,52 | B3-A5 | 89,70 | 10,30 |
| A4-B3 | 90,48 | 9,52 | B4-Gruppo A | 91,88 | 8,12 |
| A4-B4 | 90,30 | 9,70 | B4-A1 | 92,95 | 7,05 |
| A5-Gruppo B | 89,59 | 10,41 | B4-A2 | 93,54 | 6,46 |
| A5-B1 | 88,38 | 11,62 | B4-A3 | 91,57 | 8,43 |
| A5-B2 | 89,22 | 10,78 | B4-A4 | 90,30 | 9,70 |
| A5-B3 | 89,70 | 10,30 | B4-A5 | 91,05 | 8,95 |
| A5-B4 | 91,05 | 8,95 | | | |

I due gruppi presentano diversi livelli di omogeneità: nel gruppo A la differenza sull'accordo fra il codificatore più (A1) e meno in accordo (A5) è di oltre tre punti percentuali, mentre nel gruppo B la differenza fra B4 e B1 è di circa un punto. Anche le differenze fra le coppie risultano abbastanza livellate, con alcune eccezioni nel gruppo B (B1 e B3) a riprova della maggiore eterogeneità del gruppo A. Per i due codificatori che scendono sotto il 90% di accordo (A4 e A5) diventa interessante approfondire il tasso di errore. Ricordiamo infatti che il Non accordo non vuol dire che entrambi i codificatori abbiano sbagliato. Di solito uno dei due ha dato la codifica corretta e l'altro no (l'errore raramente è di entrambi, in particolare per quegli *item* che prevedono anche il punteggio parziale). Per ricostruire il quadro completo delle 709 risposte con errori di codifica è stato necessario risalire all'originale, cioè i supervisori hanno ricontrollato tutte le 709 risposte e hanno definito il codice corretto. Questo ha permesso in primo luogo la pulizia dei dati (le codifiche corrette sono state inserite nel *database* definitivo dell'indagine), e conseguentemente ha permesso di identificare il codificatore che ha sbagliato per ogni singola risposta interessata al controllo. A titolo esemplificativo presentiamo i dati relativi a due *item*, in cui è possibile avere una stima consistente dell'affidabilità di ogni codificatore. Nella Tabella 3 sono presentati i dati relativi all'*item* 1.1.

*Tabella 3. – Distribuzione degli errori fra i gruppi e i codificatori (item 1.1).*

| | Codifiche | Errori | % Errori | Tendenza |
|---|---|---|---|---|
| Tutti i gruppi | 2.470 | 86 | 3,48 | |
| Gruppo A | 1.235 | 52 | 4,21 | + |
| A1 | 249 | 3 | 1,20 | |
| A2 | 244 | 10 | 4,10 | |
| A3 | 248 | 5 | 2,02 | |
| A4 | 247 | 12 | 4,86 | - |
| A5 | 247 | 22 | 8,91 | + + |
| Gruppo B | 1.235 | 34 | 2,75 | |
| B1 | 306 | 14 | 4,58 | |
| B2 | 311 | 7 | 2,25 | + |
| B3 | 310 | 2 | 0,65 | |
| B4 | 308 | 11 | 3,57 | + |

Nota: + (*lenient*) o - (*harsh*) tendenza dell'errore superiore al 50%; ++ o -- tendenza all'errore superiore all'80%.

La percentuale di errore complessiva è del 3,48%, ed è riferita al totale delle 2.470 codifiche fatte per questo *item* nei 1.235 fascicoli, (esattamente la metà quindi del dato di Scarto riportato in Tabella 1). In questo caso non è mai avvenuto che entrambi i codificatori sbagliassero codifica. Come si può vedere il Gruppo B ha complessivamente sbagliato di meno, ma in entrambi i gruppi ci sono forti oscillazioni fra i singoli: nel gruppo A si conferma l'impressione generale vista in precedenza (con A1 e A5 agli estremi), mentre nel gruppo B il codificatore B1 ha sbagliato più di alcuni colleghi dell'altro gruppo. Per quanto riguarda il verso dell'errore, il gruppo A tende nel complesso a essere *lenient* (cioè a favore dello studente), anche se questa tendenza è dovuta soprattutto al codificatore A5, che quando sbaglia assegna molto spesso il codice 1 in luogo dello 0. Si consideri che viene segnalato con un - o un + quando il verso di un errore è più del doppio dell'altro verso, con - - o + + quando la differenza è superiore all'80%. Per questo *item* si può notare che solo un codificatore (A4) è risultato *harsh*. Nel complesso comunque solo un codificatore ha un tasso di errore superiore al 5%, cioè in un contesto PISA il gruppo dei codificatori sarebbe considerato affidabile. Nella Tabella 4 si considerano le codifiche dell'*item* 1.3. Vediamo che le *performance* individuali possono cambiare in modo significativo rispetto al precedente *item*. Ricordiamo che l'*item* 1.3 era quello che presentava il maggiore scarto fra le doppie codifiche.

*Tabella 4. – Distribuzione degli errori fra i gruppi e i codificatori (item 1.3).*

| | Codifiche | Errori | % Errori | Tendenza |
|---|---|---|---|---|
| Tutti i gruppi | 2.470 | 149 | 6,03 | |
| Gruppo A | 1.235 | 92 | 7,45 | |
| A1 | 249 | 11 | 4,42 | + |
| A2 | 244 | 27 | 11,07 | - - |
| A3 | 248 | 19 | 7,66 | + |
| A4 | 247 | 21 | 8,50 | - |
| A5 | 247 | 14 | 5,67 | |
| Gruppo B | 1.235 | 57 | 4,62 | - |
| B1 | 306 | 7 | 2,29 | - |
| B2 | 311 | 9 | 2,89 | + |
| B3 | 310 | 26 | 8,39 | - - |
| B4 | 308 | 15 | 4,87 | - - |

Nota: + (*lenient*) o - (*harsh*) tendenza dell'errore superiore al 50%; ++ o -- tendenza all'errore superiore all'80%.

Ancora una volta il Gruppo B sbaglia meno. Il dato è costante per tutti gli *item*, quindi è possibile ipotizzare un effetto collaterale del disallineamento del lavoro: il gruppo B ha codificato dopo il gruppo A, in maniera cieca, ma sempre coordinato dagli stessi supervisori, per cui l'interazione nel *forum* ha sicuramente beneficiato del lavoro svolto in precedenza nel gruppo A, con una costante discesa degli errori per tutti gli *item*, in particolare quelli a risposta articolata. Passando ai singoli vediamo che nel primo gruppo il codificatore A1 conferma la sua bravura, ma stavolta è seguito proprio da A5, che era risultato il meno affidabile nell'*item* precedente. Sbaglia invece molto il codificatore A2, per questa domanda addirittura esce dai parametri accettabili di affidabilità; in un contesto PISA questo tasso di errore comporta un ricontrollo di tutte le codifiche da lui fornite per questo *item*, ma nella nostra procedura è proprio l'incrocio delle doppie codifiche che permette di considerare accettabili le 217 codifiche in cui la codifica di A2 è stata confermata dal collega dell'altro gruppo. Per quanto riguarda il gruppo B spicca la *performance* negativa del B3, che nell'*item* precedente era stato il migliore in assoluto. La colonna della tendenza mostra un quadro molto più variegato rispetto al precedente *item*: l'unico esente da deviazioni sistematiche è A5, mentre ben 3 codificatori (A2, B3 e B4) risultano marcatamente *harsh*, e in generale tutto il gruppo B ha sbagliato molto spesso a sfavore dello studente. In un quadro complessivo di buona affidabilità esistono quindi delle propensioni particolari dei codificatori, in negativo e in positivo, sui singoli *item*. Risulta pertanto particolarmente utile il *feedback* analitico che è possibile fornire loro, per capire quali sono i tipi di risposta che possono creare maggiori difficoltà personali di codifica.

Il ricontrollo puntuale di tutte le risposte con codifiche discordanti ha permesso di migliorare la qualità dei dati e di verificare l'affidabilità dei codificatori. Ma come abbiamo visto in PISA può rimanere un dubbio sull'effettiva correttezza delle codifiche in accordo: è possibile che entrambi i codificatori sbaglino nel codificare una risposta, quindi l'errore rimarrebbe nel *database*. Seguendo l'esempio della procedura PISA è stato estratto un campione di fascicoli su cui verificare la correttezza di codifiche in accordo. Nella Tabella 5 sono riepilogati i dati relativi a quest'ultima fase.

Sono stati ricontrollate 778 risposte, poco più del 5% del totale delle risposte con codifiche uguali, rilevando in tutto 12 errori compiuti da entrambi i codificatori. Resta quindi una stima di 1,54% di errori nel *database* complessivo, un dato ampiamente accettabile, soprattutto perché non concentrato su singoli *item*, singoli codificatori, o singole classi. Ricordiamo che siamo in un contesto di domande aperte, che tutti i casi dubbi sono stati risolti, per cui quell'1,54 è un errore tollerabile in cambio di informazioni più precise fornite nel rimanente 98,46 di risposte.

*Tabella 5. – Riepilogo dei controlli sull'Accordo.*

| Item | Controllate | % Controlli | Errori | % Errori |
|---|---|---|---|---|
| 1.1 | 67 | 6,23 | 1 | 1,49 |
| 1.3 | 57 | 5,57 | 1 | 1,75 |
| 1.4 | 60 | 5,35 | 1 | 1,67 |
| 2.2 | 64 | 5,21 | 0 | 0,00 |
| 2.3 | 67 | 5,48 | 0 | 0,00 |
| 2.4 | 65 | 5,32 | 0 | 0,00 |
| 4.1 | 64 | 6,20 | 1 | 1,56 |
| 4.3 | 68 | 5,91 | 1 | 1,47 |
| 5.3 | 78 | 6,66 | 4 | 5,13 |
| 5.4 | 56 | 5,79 | 2 | 3,57 |
| 5.5 | 69 | 5,70 | 1 | 1,45 |
| 7.1 | 63 | 5,78 | 0 | 0,00 |
| Totale | 778 | 5,76 | 12 | 1,54 |
| Risposta breve | 265 | 5,43 | 1 | 0,38 |
| Risposta articolata | 513 | 5,94 | 11 | 2,14 |

In parte sorprendono i 4 errori rilevati nel controllo delle risposte 5.3, e probabilmente, nel prosieguo dell'indagine, saranno estesi i controlli su un campione più rilevante di casi, poiché è stata superata la soglia del 5% di errore. Ma anche questo è un merito della procedura, identificare e trattare in maniera specifica i casi dubbi, evitando ricontrolli a tappeto.

## 6. Conclusioni

La procedura proposta è risultata in grado di raggiungere i risultati sperati per garantire un buon livello di affidabilità delle codifiche e un efficace controllo dei codificatori coinvolti, con un indubbio aggravio di lavoro nella fase di codifica, dovuto alla doppia lettura cieca di ogni fascicoli, ma anche con un consistente risparmio nelle fasi di verifica della affidabilità e di ricontrollo dei dati. Il rapporto costi/benefici risulta quindi sostenibile, anche in contesti più ristretti quali una scuola, dove i docenti di una stessa disciplina (o disci-

pline affini) possono formare due gruppi autonomi di codifica e verificare la loro capacità di valutazione delle domande aperte, uno strumento che può fornire informazioni importanti sulle abilità degli studenti e che comincia ad apparire anche nelle rilevazioni INVALSI. La procedura presentata risulta particolarmente utile nei casi in cui gli *item* a risposta aperta siano originali e poco collaudati, quindi sempre in un contesto scolastico in cui vengono allestiti materiali specifici collegati ai percorsi didattici svolti. L'integrazione costante della *Guida per la codifica* può rendere lo strumento più affidabile in successive rilevazioni, per poter effettuare confronti fra le diverse coorti di studenti che si succedono nella scuola.

La possibilità di stimare l'incidenza degli errori nelle codifiche permette di affrontare l'analisi statistica dei dati con maggior consapevolezza. La possibilità di fornire ad ogni codificatore un *feedback* preciso sulla sua *performance* permette di migliorare la sua capacità nella codifica di risposte aperte, capacità che sarà sempre più richiesta visto il crescente utilizzo di tale strumento in rilevazioni nazionali e internazionali.

Ricordando sempre che nell'analisi delle risposte aperte i margini di interpretazione soggettiva da parte del valutatore sono sicuramente più ampi rispetto ai quesiti strutturati classici, l'adozione di procedure rigorose di codifica e di controllo delle codifiche permette di orientare anche questo tipo di quesiti semistrutturati verso l'oggettività, riducendo in modo significativo gli errori di interpretazione.

Il rapporto completo di ricerca dell'indagine *Problem solving e abilità geografiche* sarà pubblicato nei primi mesi del 2015.

## Riferimenti bibliografici

Ackerman, T. A., & Smith, P. L. (1988). A comparison of the information provided by essay, multiple-choice, and free response writing tests. *Applied Psychological Measurement*, *12*(2), 117-128.

Albertoni, D. (1988). L'addestramento dei valutatori delle prove IEA/IPS. *Ricerca Educativa*, *5*(2-3), 95-112.

Anderson, P., & Morgan, G. (2008). *Developing Tests and Questionnaires for a national assessment of educational achievement*. Washington: The World Bank. http://siteresources.worldbank.org/EDUCATION/Resources/278200-1099079877269/547664-1222888444288/National_assessment_Vol2.pdf (consulted 25/08/2014).

Asquini, G. (2006). La capacità di problem solving dei quindicenni. In M. T. Siniscalco (a cura di), *Il livello di competenza dei quindicenni italiani in mate-*

*matica, lettura, scienze e problem solving*. Rapporto nazionale di PISA 2003. Roma: Armando.

Asquini, G., & Corsini, C. (2010). L'evoluzione dei risultati in lettura nelle diverse edizioni di PISA. In AA.VV., *PISA 2006. Approfondimenti tematici e metodologici* (pp. 177-201). Roma: Armando.

Bennett, R., & Ward, W. (Eds.). (1993). *Constructing versus choice in cognitive measurement: Issues in constructed response, performance testing, and portfolio assessment*. Hillsdale, NJ: Lawrence Erlbaum Associates Publishers.

Benvenuto, G. (1993). L'affidabilità delle valutazioni. In AA.VV., *La produzione scritta nel biennio superiore* (pp. 27-34). Campobasso: IRRSAE Molise.

Bolasco, S. (2010). *TaLTaC2.10. Sviluppi, esperienze ed elementi essenziali di analisi automatica dei testi*. Milano: LED.

Domenici, G. (1996). *Gli strumenti della valutazione* (2ª ed.). Napoli: Tecnodid.

INVALSI (2010). *Rilevazione degli apprendimenti – SNV. Prime analisi*. INVALSI. http://www.invalsi.it/download/rapporti/snv2010/Rapporto_SNV_09_10.pdf (consulted 25/08/2014).

INVALSI (2014). *Rilevazioni nazionali degli apprendimenti 2013-14. Rapporto risultati*. INVALSI. http://www.invalsi.it/areaprove/rapporti/Rapporto_Rilevazioni_Nazionali_2014.pdf (consulted 25/08/2014).

Jonassen, D. H., (2011). *Learning to solve problems: A handbook for designing problem-solving learning environments*. New York: Routledge.

McGaw, B. (2008). The role of the OECD in international comparative studies of achievement. *Assessment in Education: Principles, Policy & Practice*, *12*(2), 223-243.

Morris, A. (2011). *Student standardised testing: Current practices in oecd countries and a literature review*. OECD Education Working Papers, 65. Paris: OECD.

OECD (1999). *Measuring Student knowledge and skills. A new framework for assessment*. Paris: OECD.

OECD (2003). *The PISA 2003 Assessment framework*. Paris: OECD.

OECD (2004). *Problem solving for tomorrow's world*. Paris: OECD.

OECD (2012). *PISA 2009 Technical report*. Paris: OECD.

OECD (2013). *PISA 2012 Assessment and analytical framework*. Paris: OECD.

Polya, G. (1945). *How to solve it*. Princeton, NJ: Princeton University Press.

Ryjchen, D., & Salganik, L. H. (2000). *Definition and selection of key competencies (DeSeCo)*. Paris: OECD.

Sabella, M. (2014). *Primi della classe si nasce? Indagine longitudinale sul Summer Learning Loss nella scuola secondaria di I grado*. Roma: Nuova Cultura.

Toch, T. (2006). *Margins of error: The education testing industry in the no child left behind era. Education sector reports*. http://www.educationsector.org/sites/default/files/publications/Margins_of_Error.pdf (consulted 25/08/2014).

Ward, W. C., Dupree, D., & Carlson, S. B. (1987). *A comparison of free-response and multiple-choice questions in the assessment of reading comprehension (RR-87-20)*. Princeton, NJ: Educational Testing Service.

Weber, K., Radu, I., Mueller, M., Powell, A., & Maher, C. (2010). Expanding participation in problem solving in a diverse middle school mathematics classroom, *Mathematics Education Research Journal*, *22*(1), 91-118.

## Riassunto

*Il saggio presenta un approfondimento di ricerca all'interno del progetto «Problem solving e abilità geografiche», realizzato grazie al Finanziamento di Ateneo 2011 della Sapienza, Università di Roma. Per il trattamento dei quesiti a risposta aperta, largamente utilizzati nello strumento di indagine, è stata definita una procedura specifica per la codifica e per il controllo dell'affidabilità delle codifiche e dei codificatori. Ricordando sempre che nell'analisi delle risposte aperte i margini di interpretazione soggettiva da parte del valutatore sono sicuramente più ampi rispetto ai quesiti strutturati classici, l'adozione di procedure rigorose di codifica e di controllo delle codifiche permette di orientare anche questo tipo di quesiti semistrutturati verso l'oggettività, riducendo in modo significativo gli errori di interpretazione. Il riferimento principale è alla procedura utilizzata nell'indagine OCSE-PISA per le domande aperte, ma è stata considerata anche l'esperienza di valutazione delle prove scritte definita nell'indagine IEA «Written composition». La procedura è finalizzata al miglioramento della qualità del dataset dell'indagine, con il minor aggravio di risorse possibile. I risultati ottenuti hanno confermato l'efficacia della procedura in termini di affidabilità delle codifiche, con la stima di un livello di errore di codifica assai contenuto. Inoltre è stato possibile fornire un feedback puntuale a ogni codificatore e ottenere il miglioramento dell'abilità di codifica.*

*Parole chiave:* Affidabilità, Codifica, Domande aperte, Problem solving, Valutazione.